DOCUMENT RESUME

ED 458 284 TM 033 448

AUTHOR Henderson, Dianne L.
TITLE Prevalence of Gender DIF in Mixed Format High School Exit Examinations.
PUB DATE 2001-04-12
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Biology; English; Foreign Countries; Graduation Requirements; *High School Students; High Schools; *Item Bias; Mathematics; *Sex Differences; Social Studies; Test Items
IDENTIFIERS Alberta; Canada; *Exit Examinations; Item Bias Detection; Mantel Haenszel Procedure

ABSTRACT

The primary purpose of this study was to identify potential sources of gender differential item bias (DIF) in a high school exit examination composed of both selected-response and constructed-response items in the content areas of English, social studies, mathematics, and biology. A secondary purpose was to determine the agreement between the polytomous differential item functioning (DIF) detection methods, the Generalized Mantel-Haenszel (GMH) approach and Poly-Simultaneous Item Bias (Poly-SIB), and their counterparts, the Mantel-Haenszel procedure (MH) and SIB. Data were from four different Alberta Education Diploma Examinations for June and January 1998. The numbers of students that completed each form ranged from 2,328 to 3,386. Results indicate that both GMH and Poly-SIB were comparable to their dichotomous counterparts, MH and SIB, although there were slight differences between MH and GMH. Results about gender DIF support some hypotheses and not others. Males did not outperform females on geometry and mathematical problem solving items. Although more than 50 mathematics items were analyzed, only 8 dichotomous items were flagged. None of the gridded response items were flagged for DIF, and references to stereotypical male or female activities were not identified as DIF items or did not consistently favor one group or the other. While the majority of the dichotomous items favored males, all of the polytomous items favored females. These findings suggest that there may be an item-by-format interaction where females perform better on constructed response items even in measures of quantitative ability. The paper discusses some areas for future research. (Contains 10 tables, 1 figure, and 34 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 284

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

D. Henderson

1

Prevalence of Gender DIF in Mixed Format High School Exit Examinations

Dianne L. Henderson[1]

CTB/McGraw-Hill

Paper presented at the annual meeting of the American Educational Research Association

Seattle, Washington

April 12, 2001

TM033448

BEST COPY AVAILABLE

[1] Please address all correspondence to Dianne L. Henderson, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA, 93940.

Prevalence of Gender DIF in Mixed Format High School Exit Examinations

The goal of all test developers is to assemble a set of items that provides an estimate of an examinee's ability that is as fair and accurate as possible for all groups of the population. Thus the test development process includes a systematic item analysis to ensure that all examinees with the same underlying level of ability have the same probability of getting an item correct. Unfortunately, empirical evidence can often be found in administered tests which indicates that certain subgroups of the test taking population, matched with respect to the construct being measured, have a different probability of getting the item correct. Such items are described as having differential item functioning (DIF; Dorans & Holland, 1993; Holland & Thayer, 1988).

DIF may be attributed to either item impact, item bias, or Type I error. If the item reflects actual differences in the knowledge or ability of the examinees, then DIF may be attributed to item impact (Camilli & Shepard, 1994). On the other hand, if the item is characterized by a systematic error in how an item measures the intended construct for a distinct group of examinees (e.g., Aboriginal, female), then DIF may be attributed to item bias (Camilli & Shepard, 1994). Finally, an item may be falsely identified by chance alone (Type I error). While there are many studies investigating the prevalence of DIF among dichotomous items, there are fewer studies investigating the prevalence of DIF among polytomous items or in operational tests composed of both item types (Clauser & Mazor, 1998; Downing & Haladyna, 1997; Potenza & Dorans, 1995).

Perspective/Theoretical Framework

Gender DIF

The investigation of DIF has been examined comparing several different subgroups of the population, however the majority of published articles have focused on the investigation of item-level differences between males and females. DIF studies on tests composed of dichotomous items has identified specific content areas that tend to favor one group over the other (Carlton & Harris, 1989; Doolittle, 1989; Doolittle & Cleary, 1987; Gierl & McEwen, 1998; Scheuneman & Gerritz, 1990; O'Neill & McPeek, 1993). For example, males tend to perform better than females on items related to science, and on items referring to stereotypical male activities on verbal ability measures found on standardized tests like the Graduate Record Examination (GRE; O'Neill & McPeek, 1993). In addition, males tend to perform better than females on items that involve proportions, ratios, geometry, graphs, tables, or figures (Burton, 1996, Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993). In contrast, females tend to perform better than males on items related to aesthetics, human rights, computation, and those that involve symbols (Burton, 1996, Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993; Sadker & Sadker, 1994).

In addition to the hypothesis that DIF is related to specific content, there is also some evidence to indicate that the type of item scoring may be related to DIF. For example, while males generally perform better than females on dichotomous items, females perform better than males on polytomous items like essays, possibly because of better verbal fluency, reading, and writing skills (Breland, Danos, Kahn, Kubota, & Benner, 1994; Pomplun & Sundbye, 1999; Willingham & Cole, 1997). Despite this observation, systematic investigations to determine if different types of item scoring (e.g., multiple choice, performance assessment) contribute to DIF are limited (Willingham & Cole, 1997). The examination of DIF in examinations composed of both types of items is also limited. As more standardized examinations containing both types of items are created, a better

understanding of DIF across item format and the implications that any interaction may have on test performance is required.

DIF Detection Methods

Currently, the most popular methods used to detect DIF in dichotomous items are the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) methods. The recent inclusion of constructed response items on large-scale standardized examinations has also led to the development of methods that can detect DIF among polytomous items. The Mantel procedure (Zwick, Donoghue, & Grima, 1993), and Simultaneous Item Bias for polytomous items (Poly-SIB; Chang, Mazzeo, & Roussos, 1996) have both been generalized for use with polytomous items. In the following paragraphs each of these DIF detection methods will be briefly described.

Mantel-Haenszel

One of the more commonly used methods used in DIF detection studies is the MH statistical procedure (Mantel & Haenszel, 1959). The MH DIF detection procedure uses contingency tables to compare the probability of success on each item for two groups of interest after matching on ability. In order to compare the probabilities of a correct response, item response data for the reference and focal group members are arranged into a series of 2 x 2 contingency tables, one for each score level of the item. For each item, K 2 x 2 tables are constructed, where K is the number of unique scores for the test. The MH statistic, χ^2_{MH}, is calculated from the K 2 x 2 tables for each item and is distributed approximately as a chi-square statistic with one degree of freedom.

The associated index of DIF, α_{MH}, is a constant odds ratio and is interpreted as the average factor by which the odds that an examinee from the reference group will answer the item correctly exceed the odds of an examinee from the focal group. The resulting statistic is symmetrically distributed about zero with values of zero interpreted as no DIF. Positive delta values indicate DIF favoring the focal group and negative delta values indicate DIF favoring the reference group. Guidelines for interpreting the degree of DIF in test items have been established at ETS (Zwick & Erickan, 1989). Roussos and Stout (1996) have modified these guidelines to aid in the interpretation of DIF:

- Negligible or A-level DIF: Δ_{MH} is not significantly different from 0 OR Δ_{MH} is significantly different from 0 using χ^2_{MH} AND $|\Delta_{MH}| < 1$.
- Moderate or B-level DIF: Δ_{MH} is significantly different from 0 using χ^2_{MH} AND $|\Delta_{MH}|$ at least 1 but less than1.5.
- Large or C- level DIF: Δ_{MH} is significantly different from 0 and $|\Delta_{MH}|$ is 1.5 or greater.

Generalized Mantel Haenzel

As a result of the increased use of constructed response items in standardized tests and the subsequent need to identify DIF among this tem type, extensions of the MH procedure have been proposed (Zwick, et al., 1993). To investigate DIF in items with ordered response categories, a test of conditional association proposed by Mantel (1963) has been used to compare the item means for the two groups of interest after matching on ability. In this method, item response data for the reference and focal group members are arranged into a series of 2 x T x K contingency tables, one for each item at each score level. For each K, 2 x T tables are constructed, where K is the number of levels of the matching variable and T is the number of response categories for the item (see Figure 1). The associated statistic, *Mantel* χ^2 is calculated from the K 2 x T tables for each item and is

distributed approximately as a chi-square statistic with one degree of freedom. In the dichotomous case, the Mantel (GMH) statistic is identical to the MH statistic without the continuity correction. An assumption associated with GMH is that the odds ratios are constant across item score categories. Consequently, GMH identifies an overall or global amount of DIF, but cannot identify DIF at each item score category.

To help interpret DIF magnitude for polytomous items, a ratio statistic obtained by dividing the standardized mean difference (SMD) by the item standard deviation can be used (Zwick, Thayer & Mazzeo, 1997). An extension of the STD P-DIF (Dorans & Kulick, 1986), the SMD is a descriptive index that compares the item means of the two groups after adjusting for differences in the distribution of members across the values of the matching variable. A negative SMD value implies that, conditional on the matching variable, the focal group has a lower mean item score than the reference group. While the SMD is meaningful in reference to items scored on the same scale, it cannot be used to compare across items of varying score scales (e.g, a four-point item cannot be compared to a two-point item). To obtain an effect size index that is scale invariant, the index is divided by a measure of the item score variability resulting in an index that can be compared across items (Zwick, et al., 1997). In this study the item standard deviation for the combined group of students was used in the calculation of this effect size measure.

Preliminary work has also been conducted to identify a classification system for polytomous items analogous to the system used dichotomous items (Zwick, et al., 1997). Like the dichotomous system, the guidelines for interpreting the degree of DIF in polytomous items include both the results of the Mantel hypotheses test and the effect size measure:

- Negligible or A-level DIF: *SMD* is not significantly different from 0 OR *SMD* is significantly different from 0 using *Mantel* χ^2 AND $|SMD| < 0.17$.
- Moderate or B-level DIF: *SMD* is significantly different from 0 using *Mantel* χ AND $|SMD|$ at least 0.17 but less than 0.25.
- Large or C- level DIF: *SMD* is significantly different from 0 and $|SMD|$ is 0.25 or greater.

Simultaneous Item Bias Test

Developed by Shealy and Stout (1993), SIB is a model-based approach that includes a test of significance and an explicit correction for guessing. With SIB the test items are split into two subtests: a "studied" subtest and a "matching" subtest. The studied subtest contains potential DIF items while the matching subtest, often but not always, contains the rest of the items (Li, Nandakumar, & Stout, 1995). The matching subtest contains items that measure the construct of interest and are not suspected of functioning differently. Using the total test score from the matching subtest, examinees from the reference and focal groups are matched on ability by grouping them into *K* subgroups. Then, the performance of the examinees is compared across the reference and focal groups on the studied subtest (Li et al., 1995).

The means of the studied subtest for the reference and the focal groups are adjusted to correct for any differences in the ability distributions of the two groups. If the ability distributions of the reference and focal groups are equal and the item does not contain DIF, then the difference in means on the studied subtest will equal zero. However, if there are differences in the ability distributions of the reference and focal groups, then the differences in means will not equal zero even when no DIF is present. This is due to the "incompatibility in the average scores of the matching subtest for the two groups within subgroup *K*" (Li et al., 1995, p. 7). To correct this

problem, SIB adjusts the observed score on the matching subtest by estimating the true score for the reference and focal groups at the *K* subgroup level. To calculate this the

> equation for the linear regression of true score on observed score from Classical Test Theory [is used] with KR20 calculated as the slope of the regression line for each group…Then the average of these two scores is calculated….The corresponding adjusted mean scores on the studied subtest, $\bar{Y}_{Rk}^{*}$ and $\bar{Y}_{Fk}^{*}$, are obtained using a first order Taylor Series approach to adjust for focal and reference group differences in the estimated true scores for subgroup [score level] *K* (Li et al., 1995, p. 8).

As with GMH and MH, SIB provides a statistic for testing the null hypothesis of no DIF, $\hat{\beta}_U$. The interpret the amount of estimated DIF, Roussos and Stout (1996, p. 220) suggest the following guidelines:

- Negligible or A-level DIF: $\hat{\beta}_U < 0.059$
- Moderate or B-level DIF: null hypothesis is rejected and $0.059 \leq \hat{\beta}_U < 0.088$.
- Large or C-level DIF: null hypothesis is rejected and $\hat{\beta}_U \geq 0.088$.

Poly-SIBTEST

SIB has also been generalized for use with polytomous items (Poly-SIB; Chang et al., 1996). Like GMH, while Poly-SIB calculates a global amount of DIF for each item, it does not provide an indication of DIF at the individual score categories (A. G. Froelich, personal communication, July 6, 1999). In addition, Poly-SIB can be used with examinations containing both dichotomous and polytomous items. However, unlike GMH, the associated effect size measure, $\hat{\beta}_U$, is interpreted using the same guidelines as SIB for both dichotomous and polytomous items.

Purpose

The primary purpose of this study was to identify potential sources of gender DIF in high school exit examination composed of both selected-response and constructed-response items in the content areas of English, Social Studies, Mathematics, and Biology. A secondary purpose of this study was to determine the agreement between the polytomous DIF detection methods, GMH and Poly-SIB and their respective counterparts, MH and SIB.

Method

To answer the questions posed for this study, data from four different Alberta Education Diploma Examinations were analyzed. The examinations included English, Social Studies, Mathematics and Biology. In all cases the exams included both dichotomous and polytomous items. Samples of examinees that completed the June, 1998 forms of the four examinations were analyzed to determine the prevalence of gender DIF across item format and subject area. This data, as well as data from the January, 1998 forms were also used to help determine if the polytomous versions of the common DIF detection methods performed similarly to their dichotomous counterparts and to determine the comparability of the polytomous DIF detection methods.

The numbers of students that completed each form ranged from 2328 to 3286 (see Table 1). In general the samples included more female than male students. The smallest female sample (n =

946) was noted in Mathematics (June). The smallest males sample (n = 987) was noted in English (January).

To compare the polytomous DIF detection methods to their dichotomous counterparts (MH and GMH, SIB and Poly-SIB) the dichotomous items included in each of the four examinations were used. Specific computer programs were used for MH (Shealy & Stout, 1993), SIB (Shealy & Stout, 1993), and Poly-SIB (Chang et al., 1996). SAS was used for GMH. All analyses were conducted using the same matching variable: the total test score associated with items analyzed. Identified DIF items were classified into three categories: those that exhibited no DIF or a negligible amount of DIF (A-level) and those that exhibited moderate (B-level) or large levels of DIF (C-level) based on the associated effect size measure and classification schemes for each method described earlier.

The polytomous DIF detection methods were then compared in a manner similar to the analyses described above except that the entire set of test items were analyzed. Like the previous analyses, the matching variable was the total test score obtained over all of the items. Identified DIF items were classified into the three categories as described above using the guidelines associated with the effect size measure for each method.

Following the comparison of the DIF detection methods, items that were identified as exhibiting moderate or large DIF on the June forms were identified and classified according the associated test blueprints for each content area. Item characteristics were then summarized and compared to previously reported findings of DIF and hypothesized differences between males and females.

Examinations to be Studied

The Department of Education in Alberta, Canada developed the examinations used in this study. Exit examinations have been used in Alberta since 1984 to: 1) ensure the maintenance of provincial standards of achievement; 2) certify the level of individual student achievement in selected Grade 12 courses, and 3) report individual and group results (Alberta Education, 1998a). The results from these examinations account for 50% of the total awarded mark in each subject for which an examination is available. Questions on the exams are based on concepts, topics, and facts from the Alberta Education Program of Studies that are to be included in the curriculum for all students in the Province of Alberta. These examinations are administered four times a year, however, the majority of students complete them in either January or June. Each examination is subjected to a sensitivity review and content analysis prior to administration, however, no statistical analyses of DIF are completed.

The four examinations selected for this study represented both the humanities and the sciences: English, Social Studies, Mathematics, and Biology. Each Diploma Examination contains a mixture of dichotomous and polytomous items (see Table 2). The number of dichotomous items ranges from 49 items in the Mathematics examination to 70 items in the English examination. Dichotomous items in the English and Social Studies examination consist only of multiple choice items. Dichotomous items in the Mathematics and Biology examinations also include 9 and 8 numerical response items in which the students "grid" in their answers, respectively. While the numerical response items in Mathematics involve calculations, in Biology, these items are also used to record answers to matching, fill-in-the blank, diagram labeling, and ordering a sequence of events.

The dichotomous items are also described and classified according to the examination blueprint provided by Alberta Education. For English and Social Studies, course content and cognitive domain are used to classify these items. For Mathematics, course content and level of

mathematical understanding classify the items. For Biology, items are classified by course content alone.

The number of polytomous items varied from 2 to 6 items. The fewest number of items were on the Biology examination, the largest on the English examination. The item types and associated scoring method included on each examination also varied. In English the items were rating scales associated with two essays. The first two items were scales associated with a minor essay; the remaining four items were associated with a major essay. In Social Studies, the six items were also rating scales, but they were associated with only one essay assignment. In both examinations the points on the rating scales ranged from 1 to 5. The mark assigned was the combined average of the scores awarded by two independent raters. In the case of imperfect agreement between the raters, scores were assigned that fell mid-way between the ratings made. For example, if rater A assigned a 2 and rater B assigned a 1, then a score of 1.5 was awarded (Alberta Education, 1998a, 1998b). To maintain the integrity of the five-point rating scale for data analysis, those scores that fell between the scale scores were randomly recoded to the nearest score category. For example, all scores of 1.5 were randomly recoded to either 1.0 or 2.0.

The three polytomous mathematics items are complex, multi-step problems. The total possible mark for each item was four. Specific scoring rubrics were used to assign a partial or full mark based on the degree of successful completion of the problem. In Biology, two polytomous items were included on the examination. However, only one item was included in this study. The included item required students to evaluate data, incorporate previous knowledge with new information, form new hypotheses, and make predictions regarding future trends. This item was scored holistically on a scale of 1 to 4. The deleted item was the sum of four related items totaling 12 possible marks. As this scale was too large for the DIF detection procedures used in this study and because the individual item scores were unavailable, this item was deleted.

Results

Comparability between Dichotomous and Polytomous DIF Detection Methods

The comparability between the dichotomous and polytomous versions of the DIF detection methods was completed with the samples of students who wrote the examinations in the 1998 academic year. Both the January and June administrations were used for this part of the study. Only the dichotomous items from each examination were included for this part of the study. Students were matched based on their performance on this section of the examination. That is, the matching variable was the simple sum of the dichotomous items and not the total test score.

The results of the comparisons between the dichotomous and polytomous versions of the DIF detection methods (MH/GMH and SIB/Poly-SIB) are summarized in Table 3 for each subject area and sample. The summary table is organized in terms of lower triangles. The off-diagonal elements are those items with moderate to severe DIF that were identified by both procedures. The diagonal elements are items identified solely by one, but not both methods. Where zeros are located in the diagonal elements, the agreement is 100%. The overall agreement, in percent, for each pair of methods for each examination and sample are also included in this table. This calculation included all items, including items identified with negligible or no DIF.

The agreement between each set of DIF detection methods is good. Of the eight analyses conducted (four content areas, 2 administrations), the agreement between the two methods ranged from 94.3% to 100%. The agreement between SIB and Poly-SIB was slightly better than the agreement between MH and GMH, with seven of the eight analyses demonstrating 100% agreement.

As both Poly-SIB and the GMH procedures are generalizations of their dichotomous counterparts, it was expected that the same items would be statistically flagged regardless if the polytomous or dichotomous procedures were implemented. In addition, unlike the GMH and the MH procedures, the effect size measure for Poly-SIB is also a generalization of the SIB DIF detection method. Therefore it was expected that the same items would be flagged between SIB and Poly-SIB. In the single case of non-agreement between these two procedures over the eight different analyses, the discrepancy is attributable to the arbitrary nature of the established guidelines used to determine DIF magnitude because different decisions can result when effect sizes are "centered" around a specific cut-point. That is, in the discrepant case, the SIB effect size was 0.059, whereas the effect size measure for Poly-SIB was 0.058. Based on these results, the item was identified as displaying moderate DIF with SIB but negligible DIF with Poly-SIB even though the difference between the two effect sizes is only 0.001. However, unlike SIB and Poly-SIB, the effect size measure for the GMH procedure is not a generalization of the effect size measure used with MH. This is the most likely explanation for the differences between the number and types of items flagged by each procedure.

Comparisons between Polytomous DIF Detection Methods

To investigate the behavior of GMH and Poly-SIB in mixed format examinations, the analyses reported above were repeated using both the dichotomous and polytomous items for each examination. As in the previous analyses, the total test score was used to match examinees on ability. The analyses were completed on the same data sets as described in the previous section.

A summary of the number of DIF items detected by the GMH and Poly-SIB procedures across all items is provided in Table 4 for each of the four examinations. As before, a lower triangle is provided for each examination and sample. Two numbers, separated by a comma, are presented in each cell. The first number corresponds to the number of dichotomous items identified with moderate or severe DIF; the second number corresponds to the number of polytomous items identified. The diagonal elements represent the number of dichotomous and polytomous items identified by each method. The numbers in the off diagonals represent the number of dichotomous and polytomous items commonly identified by the other method. For example, in January English, GMH identified 11 dichotomous and no polytomous DIF items and Poly-SIB identified 17 dichotomous and 2 polytomous DIF items. Of the 11 dichotomous DIF items identified by GMH, Poly-SIB identified all 11 items.

In all cases, Poly-SIB identified more dichotomous and polytomous items than GMH. Further, Poly-SIB identified the majority of DIF items also identified by GMH. In one case, GMH identified a dichotmous item that was not flagged by Poly-SIB (Social Studies, June). However, it should be noted that this item was not flagged by Poly-SIB due to the arbitrary nature of the established cut-point used to separate negligible ($\hat{\beta}_t < 0.059$) from moderate DIF ($0.059 \leq \hat{\beta}_t < 0.088$). In this situation, the Poly-SIB effect size was 0.058; therefore the item was classified as displaying negligible DIF even though it was within 0.001 of being classified as moderate DIF.

Comparison of the dichotomous results (first number in each pair in Table 3) with the dichotomous results reported in Table 4 reveals that the inclusion of the polytomous items and the resulting change in the matching variable altered the number of dichotomous items detected. In some cases more dichotomous items were identified, in other cases fewer or the same number of dichotomous items were identified. These differences are most likely related to the inclusion of the polytomous items, which altered the total test score (matching variable) and the ability distribution of the reference and focal groups. Consequently, in some cases, the use of a different matching

variable in the analysis of the combined set of items produced different results for the dichotomous items than when the dichotomous items were analyzed alone.

Gender DIF Across Item Format and Subject

Following the comparison of polytomous DIF detection methods, items that were identified as exhibiting moderate or large DIF on the June examinations were identified and then classified according the associated test blueprints for each content area.

The DIF detection method Poly-SIB was used as this method identified the most items and, with the exception of one item, it identified all of the items flagged by the GMH procedure. It could be argued that the increased numbers of items identified by Poly-SIB may be related to Type I error; however it could also be argued that that fewer DIF items would remain undetected with Poly-SIB (Type II error). While a more conservative method may be more desirable to test developers and administrators, this may not be most desirable by examinees and social advocates, especially if DIF items remain undetected. As the primary purpose of this part of the study was to explore the prevalence of DIF and the characteristics of those items, the most liberal DIF detection method was selected. Prior to discussing the characteristics associated with the flagged items, overall test level differences will be discussed.

Although differences between the mean scores for males and females are not adequate to identify DIF, they are commonly reported and referenced as evidence of differential performance or impact. Further, means, together with standard deviations, provide a description of the overall performance of the groups to be studied. Hence, the mean and standard deviation for the total score, the dichotomous items, and the polytomous items for the males and females that completed the June examinations are presented in Table 5, together with the effect sizes and t-test results. Significant differences between the two groups were noted on various aspects of the four examinations. To interpret the differences between the means of the males and females, effect sizes were computed using the standard deviation of the males as an estimate of the variance. These effect sizes were interpreted using Cohen's (1988) operational definitions for small (d = .2), medium (d = .5), and large (d = .8).

In the English examination, females had significantly higher mean scores ($p < .05$) on the overall test scores, as well as scores on the polytomous and dichotomous sections. However, the effect sizes associated with the differences are all small to medium ($d \leq -.30$). In contrast to English, males had a significantly ($p < .05$) higher mean score than females on the dichotomous section and the total test score for Social Studies. The associated effect sizes for these differences were .45 and .34 respectively. In contrast, females had a significantly ($p < .05$) higher mean score on the polytomous section of the examination, however the effect size was small ($d = -.13$). This pattern was also observed in Mathematics where the males had a significantly ($p < .05$) higher mean score than females on the dichotomous section of the test ($d = .14$) and the total test score ($d = .08$), and the females had a significantly ($p < .05$) higher mean score on the polytomous section ($d = -.08$). Unlike the other three subject areas, there were no significant differences between the males and females in Biology on either section of the test or the total test score.

While differences between mean scores indicate differential performance over the associated items, such differences do not necessarily imply the presence of differential performance at the item level. To make that determination it is necessary to conduct DIF analyses. The results of the prevalence of DIF across item format and subject area is presented in Table 6. The prevalence of DIF across item format is discussed separately for each of the four subject areas. These results are then summarized across the four subject areas to address the question of potential interactions among

subject area, gender, and item format.

For each examination, the dichotomous items identified with DIF are described and classified according to the examination blueprint provided by Alberta Education. For English and Social Studies, course content and cognitive domain are used to classify these items. For Mathematics, the items are classified by course content and level of mathematical understanding. For Biology, items are classified by course content alone. Although these descriptions are useful to describe the types of items with DIF, as with previous studies, no attempt has been made to determine if the DIF is attributable to bias, impact, or Type I error. In previous studies attempts have been made to clarify the nature of DIF using panels of content experts. However, these panels are generally unsuccessful at both interpreting or predicting items that perform differently across different groups of examinees (Camilli & Shepard, 1994, Gierl & McEwen, 1988).

English

The English examination consisted of 70 dichotomous and 6 polytomous items. According to the examination blueprint, the 70 dichotomous items were "classified in two ways: according to the curricular content area being tested and according to the thinking (process) skill required to answer the question" (Alberta Education, 1998b, p. 4). As shown in Table 7, for course content, 32 were classified "Meanings", 23 were classified "Critical Response", and 15 were classified "Human Experience and Values". For thinking skills, 45 were classified as "Inference and Application"; 19 were classified "Evaluation", and 6 were classified "Literal Understanding". The polytomous items were related to two writing assignments designed to assess reading, writing, and thinking skills. Five point scales are used to score the student responses. Two items were associated with a short assignment and are labeled: "thought and detail" and "writing skills". The four remaining items were associated with a longer assignment requiring "the synthesis and ability to communicate regarding techniques used in the literary works studied in class" (Alberta Education, 1998b, p. 2). These were labeled: "thought and detail", "organization", "matters of choice", and "matters of correctness". The number of dichotomous and polytomous items identified with DIF are discussed below and described according to the preceding classifications.

In total ten items were flagged for DIF. All but one of the dichotomous DIF items were found in the content area of "Meanings", with five items favoring males and three items favoring females. Two of the items favoring males was from the thinking (process) level " Inference and Application" and two were from "Literal Understanding" and one was from "Evaluation". All three of the items favoring females in the content area "Meaning" were from the thinking (process) level "Inference and Application". One item from the content area "Human Experience and Values" also favored females. This item was classified as "Inference and Application". The polytomous item flagged for DIF favored females and was associated with the rating scale "Matters of Correctness" for the longer writing assignment. This writing scale rates sentence construction and mechanics, accurate word usage, and grammar.

Social Studies

The Social Studies examination consisted of 70 dichotomous and 4 polytomous items. According to the examination blueprint, each dichotomous item is "classified in two ways: according to the curricular content area (topic) being tested and by the knowledge and skill objectives required to answer the question" (Alberta Education, 1998c, p. 4). As shown in Table 8, the 70 dichotomous items were equally distributed between the two content areas "Political and Economic Systems" and "Global Interaction in the 20th Century". For the knowledge and skill objective, 24 items were

classified "Comprehension of Information and Ideas"; 23 were classified "Interpretation and Analysis of Information and Ideas", and 23 were classified "Synthesis and Evaluation of Information and Ideas".

The four polytomous items of the Social Studies Examination were related to one writing assignment in which the student was required to "discuss the importance and complexity of an issue and rationally defend their position by using supportive and relevant evidence" (Alberta Education, 1998c, p. 6). Each item corresponds to a five-point scoring scale. While one of the scales assessed writing skills (quality of language expression), the other scales (exploration of the issue, defense of position, and quality of examples) assessed the ability of the student to demonstrate an understanding of course content and critical thinking skills (Alberta Education, 1998c).

Twelve dichotomous items favored males, while only three favored females. Of the 12 items favoring males, three were from the content area "Political and Economic Systems" and nine were from "Global Interaction in the 20th Century". At the knowledge and skill levels, three items were from "Comprehension", five items were from "Interpretation and Analysis", and four items were from "Synthesis and Evaluation". Two of the three items favoring females were from "Political and Economic Systems". One was classified as "Interpretation and Analysis", the other was classified as "Synthesis and Evaluation". The remaining item favoring females was from "Global Interaction in the 20th Century" and was also classified as "Synthesis and Evaluation". All four polytomous scales favored females with a greater proportion of males receiving the lower scores, 1 and 2 on all scales.

Mathematics

The Mathematics examination consisted of 49 dichotomous and 3 polytomous items. Of the 49 dichotomous items, 40 were multiple-choice items and 9 were gridded numerical response items. According to the examination blueprint for mathematics and as shown in Table 9, dichotomous items are classified by one of nine unit topics or content domains and by mathematical understanding. The number of items varied across the levels of mathematical understanding ("Procedural" "Conceptual", and "Problem-Solving"). The three polytomous items included in each examination "assess whether or not students can draw on their mathematical experiences to solve problems and to explain mathematical concepts" (Alberta Education, 1998d, p. 5). These items may cross more than one unit or may require students to make connections among mathematical concepts. Specific five-point scoring rubrics are used to evaluate the quality and completeness of the student responses to each of these items.

In total, eight items were identified with DIF. Six of the eight dichotomous items favored males, while the two polytomous items favored females. Three of the six dichotomous items favoring males were from the content area "Sequences and Series". One item was classified under each of the levels of mathematical understanding. The remaining three items were from the content areas "Polynomial Functions", "Exponential and Logarithmic Functions", and "Permutations and Combinations". All were classified as "Conceptual" understanding. Both dichotomous items favoring females were classified as "Problem Solving". One item was from the content areas "Polynomial Functions", the other item was from "Statistics". No DIF items were detected in the units "Trigonomic Functions" or "Quadratic Relations".

Of the two polytomous items that favored females, one involved quadratic relations and the application of these principles, the other was related to logarithmic functions. Both items required students to apply mathematical knowledge and problem solving techniques, to solve and justify and explain the relevance of the solutions.

Biology

The Biology examination consisted of 48 multiple choice items, 8 gridded response items, and 1 polytomous item. While the gridded items in Mathematics involve calculations to obtain a specific numerical response, in Biology, these items were used to record answers to matching, fill-in-the blank, diagram labeling, and ordering sequences of events items. According to the examination blueprint, each dichotomous item is classified by general learner expectations or unit topic (Alberta Education, 1998e, p. 2). Unlike English, Social Studies, and Mathematics, there is no classification in Biology by level of thinking. The number of items in each unit topic is listed in Table 10. As shown, the number of items varies across the topics. The polytomous items required students to evaluate data, to incorporate previous knowledge with new information to form new hypotheses and to make predictions regarding future trends. Student responses were scored holistically using a four-point scale.

Of the three dichotomous items favoring males, one item was from each of the "Molecular Genetics", "Nervous and Endocrine System", and "Cell Division and Mendelian Genetics". Of the two multiple choice items favoring females, one item was from the unit "Nervous and Endocrine System" and one was from "Differentiation and Development" units. The one polytomous item analyzed favored females and required knowledge from the content areas "Reproductive Systems and Hormones", "Differentiation and Development" and "Cell Division and Mendelian Genetics" to complete the essay-type question.

Prevalence of DIF across Subject Area and Item Format

Reviewing the results presented in Table 6, the prevalence of DIF in the dichotomous sections of the examinations analyzed is similar to prevalence rates found in other measures of high school achievement in which 15% to 25% of items were identified with DIF (Hambleton et al., 1993). The only subject that had less than 15% of the dichotomous items flagged was Biology where 8.9% of items were flagged. Of the dichotomous items identified, more items favoring males than females were noted across the four examinations. While the numbers of items flagged are small (37 out of 245 items analyzed), 15 more items favored males than females. The incidence of DIF items favoring males was greatest in Social Studies, where the most items were flagged. In English and Biology, the numbers of DIF items favoring females and males were similar.

Although DIF studies are not routinely conducted on these examinations, they are carefully constructed and screened for potential sources of bias. This process, implemented for all high stakes testing, is designed to remove obvious sources of potential item bias, therefore, the number of items flagged for DIF is expected to be low. In general, the items selected for operational examinations have been carefully screened and are the best available items that have survived field testing and rigorous statistical and content reviews. These factors likely contributed to the levels of DIF among dichotomous items that were observed in this study.

In contrast to the dichotomous items, larger proportions of polytomous items were flagged for DIF. The fewest number of items flagged was in English (1 of 6). All of the items in Social Studies and Biology were flagged. In addition, two of the three polytomous items in Mathematics were flagged. In all cases these items exhibited large or C-level DIF and were identified as favoring females. While it is possible that some of these items might be flagged by chance alone, it is also possible that these items were flagged because the items reflect actual differences in the ability of the students. This would suggest an item format by gender interaction where females perform better on constructed response items, regardless of content area. While the numbers of studies investigating

DIF in mixed format examinations is limited, these findings support other research hypothesizing that females outperform males on constructed response items due to stronger writing skills, or neater, more complete answers (Lane, Wang, & Magon, 1996; Mazzeo, Schmitt, & Bleistein, 1993; Sadker & Sadker, 1994; Willingham & Cole, 1997). If such a gender by item format interaction exists, this suggests that educators need to target intervention programs designed to assist males in improving the skills required to answer this type of item. However, before recommendations can be made, additional research is required to determine if this type of interaction is limited only to these sets of examinations and these samples of students.

Summary and Discussion

The results of the comparison of DIF detection methods indicate that both GMH and Poly-SIB were comparable to their dichotomous counterparts, MH and SIB. Although slight differences between MH and GMH may suggest further research involving the comparability of DIF indexes. In particular, the use of different measures of test score variability in the denominator of the effect size measure associated with GMH may produce results that are more comparable to MH and it's associated DIF index.

In the comparison of GMH and Poly-SIB, different results were obtained for the set of dichotomous items compared to the set that included the polytomous items due to the alteration of the total test score (matching variable) and the resulting ability distribution of the reference and focal groups. The patterns of results, however, were similar regardless of the matching variable. In all cases GMH was the most conservative method, flagging the fewest items. While Poly-SIB detected the most items, the results included the majority of items also flagged by GMH.

Results of this study investigating the prevalence of gender DIF across subject area and item format support some of the previous research and hypotheses regarding differential performance between males and females, whereas other hypotheses are not supported. First, previous findings suggesting that males outperform females on geometry and mathematical problem solving items (O'Neill & McPeek, 1993) were not found in this study. Indeed, two of the three constructed response items requiring problem solving skills favored females rather than males.

Second, in this study, although over 50 different mathematics items were analyzed for DIF, only eight dichotomous items were flagged. This study found that items containing graphs, figures, or tables did not necessarily favor males as has been previously hypothesized (Burton, 1996; Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993). Similarly, mathematics items containing formulas, equations, or symbols did not necessarily favor females (Burton, 1996; Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993).

Third, while previous studies have found inconsistent results with no clear patterns of DIF favoring one group or the other (Burton, 1996), in this study, no gridded response items were flagged for DIF. Gridded response items were found on both the Mathematics and Biology examinations. In Mathematics, these items involved calculations to obtain a specific numerical response; whereas, in Biology, these items were used to record answers to matching, fill-in-the blank, diagram labeling, and the ordering of events..

Fourth, references to stereotypical male or female activities (O'Neill & McPeek, 1993) were either not identified as DIF items or did not consistently favor one group or the other. For example, a set of items on the English examination was based on a narrative passage about a father, his sons, and their experience with fly-fishing. Of the eight items relating to this passage, only one of the items was identified with DIF favoring males. It should also be noted that, in the majority of cases,

the items found on the examinations studied did not refer to stereotypical activities of either group. Further, reading passages, problems, and questions that contained references to people had names that were either gender neutral or included both a male and a female name.

Fifth, while the majority of dichotomous items favored males, all of the polytomous items flagged favored females. These findings suggest that there may be an item by format interaction where females perform better on constructed response items even in measures of quantitative ability as described in the literature (Bolger & Kellaghan, 1990; Lane et al., 1996). While the underlying reasons for these differences are not known, these differences may be related to stronger writing skills, or neater, more complete answers provided by females (Lane et al., 1996; Mazzeo, Schmitt, & Bleistein, 1993; Willingham & Cole, 1997). These findings suggest that further studies investigating the prevalence of DIF in mixed format examinations across a variety of samples and testing programs is required. If the findings in this study are not limited to this particular set of examinations and samples, then this also suggests that additional research is required to determine the underlying causes of this DIF and to develop appropriate intervention and remedial programs.

References

Alberta Education (1998a). Alberta Education Annual Report 1997-98. Edmonton, AB: Author.

Alberta Education (1998b). English 30 Diploma Examination Results Examiners' Report for June 1998. Edmonton, AB: Author.

Alberta Education (1998c). Social Studies 30 Diploma Examination Results Examiners' Report for June 1998. Edmonton, AB: Author.

Alberta Education (1998d). Mathematics 30 Diploma Examination Results Examiners' Report for June 1998. Edmonton, AB: Author.

Alberta Education (1998e). Biology 30 Diploma Examination Results Examiners' Report for June 1998. Edmonton, AB: Author.

Breland, H. & Danos, D., Kahn, H., Kubota, M., & Bonner, M., (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement History Examination. Journal of Educational Measurement, 31, 275-393.

Burton, N. M. (1996). How have changes in the SAT affected women's math scores? Educational Measurement: Issues and Practice, 15(4), 5-9.

Camilli, G. & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.

Carlton, S. T., & Harris, A. M. (1989, March). Female/male performance differences on the SAT: Causes and correlates. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIB procedure. Journal of Educational Measurement, 33, 333-353.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: An NCME instructional module. Educational Measurement: Issues and Practice, 17(1), 31-44.

Doolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. Applied Measurement in Education, 2 161-177.

Doolittle, A. E. & Cleary, T. A. (1987). Gender differences in performance on mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Dorans, N., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. Applied Measurement in Education, 10, 61-82.

Gierl, M. J. & McEwen, N. (1998, May). Differential item functioning on the Alberta Education Social Studies 30 examination. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, ON.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. Applied Measurement in Education, 6, 137-151.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.) Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Lane, S., Wang, N., & Magon, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. Educational Researcher, 15, 21-27, 31.

Li, H-H., Nandakumar, R., & Stout, W. (1995, April). Application of SIB in dealing with issues of DIF in the context of multidimensional data. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations (College Board Report No. 97-2). New York, NY: College Entrance Examination Board.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Pomplun, M., & Sundbye, N. (1991). Gender differences in constructed response reading items. Applied Measurement in Education, 12, 95-109.

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19, 23-37.

Roussos, L., & Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIB and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.

Sadker, M. & Sadker, D. (1994). Failing at fairness. How our schools cheat girls. Toronto, ON: Simon & Schuster.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 24, 97-118.

Shealy, R., & Stout, W. F. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression. Journal of Educational Measurement, 27, 361-370.

Wightman, L. F. (1998). An examination of sex differences in LSAT scores from the perspective of social consequences. Applied Measurement in Education, 11, 255-277.

Willingham, W. W. & Cole, N. S. (1997). Fairness issues in test design and use. In W.W. Willingham & N. S. Cole (Eds.), Gender and Fair Assessment (pp. 227 - 346). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30, 233-351.

Zwick, R., & Erickan, K. (1989). Analyses of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 53-66.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997, May) . Describing and categorizing DIF in polytomous items (ETS Research Report 97-05). Princeton: Educational Testing Services.

Table 1.

Sample Sizes by Gender, Content Area , and Administration

	January			June		
	Males	Females	Total	Males	Females	Total
English	987	1408	2395	1276	1883	3159
Social Studies	1936	1342	2278	1327	1949	3286
Mathematics	1317	1909	3226	1382	946	2328
Biology	1020	1557	2577	1243	1734	2997

Table 2.

Structure Comparison of Examinations

EXAM	Number of Items		Percentage of Mark	
	Dichotomous	Polytomous	Dichotomous	Polytomous
English	70	6	50	50
Social Studies	70	4	50	50
Mathematics	49*	3	70	30
Biology	56**	2	70	30

Notes: *Includes 9 numerical responses. **Includes 8 numerical responses

Table 3.

Comparison of Dichotomous and Polytomous DIF Detection Methods

		MH/GMH								SIB/POLY-SIB							
		English (i=70)		Social (i=70)		Math (i=49)		Biology (I=56)		English (I=70)		Social (i=70)		Math (i=49)		Biology (i=56)	
		D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
Jan	D	2		0		0		1		0		0		0		0	
	P	8	0	6	0	5	0	4	0	17	0	9	0	6	0	7	0
	N	2395		2178		3226		2577		2395		2178		3226		2577	
Agreement		97.1%		100%		100%		98.2%		100%		100%		100%		100%	
		D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
June	D	1		4		0		3		0		0		1		0	
	P	4	0	6	2	5	0	1	0	9	1	17	0	7	0	3	0
	N	3159		3276		2328		2977		3159		3276		2328		2977	
Agreement		98.9%		94.3%		100%		94.6%		98.6%		100%		97.9%		100%	

Notes. Where zeros are noted in the diagonals, the agreement between two methods is 100%; D = Dichotomous version;
P = Polytomous version, *i* = the number of items in the analysis.

Table 4.

Agreement between Polytomous DIF Detection Methods for all Items Combined

		English (*i* = 70, 4)		Social (*i* = 70, 6)		Math (*i* = 49, 3)		Biology (*i* = 56, 1)	
		GMH	PSIB	GMH	PSIB	GMH	PSIB	GMH	PSIB
Jan	GMH	11, 0		5, 0		5, 0		3,0	
	PSIB	11, 0	17, 2	5, 0	11, 3	5, 0	6,2	3, 0	7, 1
		GMH	PSIB	GMH	PSIB	GMH	PSIB	GMH	PSIB
June	GMH	4, 0		12, 4		4, 0		0, 1	
	PSIB	4, 0	9, 1	11 4	15, 4	4, 0	8, 2	0, 1	6, 1

Notes. The number of dichotomous and polytomous items with DIF are presented in pairs. The first number is the number of dichotomous DIF items; the second number is the number of polytomous DIF items.

Table 5.

Descriptive Statistics for each Examination by Gender

	Total Score				Dichotomous Items Only				Polytomous Items Only			
	M	SD	t	d	M	SD	t	d	M	SD	t	d
English												
M	67.13	11.63	-6.44*	-0.24	47.19	8.98	-5.09*	-0.19	19.94	3.96	-7.53*	-0.28
F	69.88	11.84			48.87	9.22			21.00	3.84		
Social Studies												
M	63.03	11.20	9.07*	0.34	50.91	9.48	11.64*	0.45	12.12	2.97	-3.71*	-0.13
F	59.19	12.37			46.69	10.66			12.50	2.85		
Mathematics												
M	40.26	10.31	2.04*	0.08	31.83	7.89	3.50*	0.14	8.43	3.16	-2.12*	-0.08
F	39.41	9.70			30.70	7.45			8.70	2.93		
Biology												
M	52.39	10.58	-0.78	-0.03	36.23	6.95	-0.87	-0.03	16.16	4.39	-0.52	-0.02
F	52.70	10.59			36.45	6.80			16.25	4.48		

Note. * $p < .05$

Table 6.

DIF Items Identified by Subject Area and Item Format

		Sample Size	Dichotomous	Polytomous
English			items = 70	items = 6
	Males	1276	5	0
	Females	1883	4	1
Social Studies			items = 70	items = 4
	Males	1327	12	0
	Females	1949	3	4
Mathematics			items = 49	items = 3
	Males	1382	6	0
	Females	946	2	2
Biology			items = 56	items = 1
	Males	1243	3	0
	Females	1734	2	1

Table 7.

English DIF Items by Course Content and Cognitive Level: Dichotomous Items

Curricular Content	Items	Literal Understanding (n_i = 6)		Inference and Application (n_i = 45)		Evaluation (n_i = 19)		TOTALS	
		Male	Female	Male	Female	Male	Female	Male	Female
Meanings	32	2 (6)	0 (6)	2 (19)	3 (19)	1 (7)	0 (7)	5	3
Critical Response	23	0 (0)	0 (0)	0 (17)	0 (17)	0 (6)	0 (6)	0	0
Human Experience & Values	15	0 (0)	0 (0)	0 (9)	1(9)	0 (6)	0 (6)	0	1
TOTALS	**70**	**2**	**0**	**2**	**3**	**1**	**0**	**5**	**4**

Notes. The numbers in parentheses are the total numbers of items within the cell classified by thinking (process) skills and curricular content. The totals are repeated in both the male and female columns. N_i = number of items.

Table 8.

Social Studies DIF Items by Course Content and Cognitive Level: Dichotomous Items

Curricular Content	Items	Knowledge and Skill Objectives: Comprehension (n_i = 24)		Interpretation and Analysis (n_i = 23		Synthesis and Evaluation (n_i = 23)		TOTALS	
		Male	Female	Male	Female	Male	Female	Male	Female
Political & Economic Systems	35	0(12)	0(12)	1(11)	1 (11)	2(12)	1 (12)	3	2
Global Interaction	35	3 (12)	0 (12)	4 (11)	0(11)	2 (12)	1(12)	9	1
TOTALS	**70**	**3**	**0**	**5**	**1**	**4**	**2**	**12**	**3**

Notes. The numbers in parentheses are the total numbers of items within the cell classified by thinking (process) skills and curricular content. The totals are repeated in both the male and female columns. n_i = number of items.

Table 9.

Mathematics DIF Items Organized by Unit and Understanding: Dichotomous Items

Unit	Items	Mathematical Understanding: Procedural (n_i = 15)		Conceptual (n_i = 19)		Problem Solving (n_i = 15)		TOTALS	
		Male	Female	Male	Female	Male	Female	Male	Female
Polynomial Functions	8	-	-	1 (4)	0 (4)	0 (4)	1(4)	1	1
Trigonometric & Circular Functions	8	0 (2)	0 (2)	0(3)	0 (3)	0 (3)	0 (3)	0	0
Statistics	4	0 (2)	0 (2)	0 (1)	0 (1)	0 (1)	1(1)	0	1
Quadratic Relations	7	0 (1)	0 (1)	0 (5)	0 (5)	0 (1)	0 (1)	0	0
Exponential & Logarithmic Functions	8	0 (4)	0 (4)	1(2)	0 (2)	0 (2)	0 (2)	1	0
Permutations & Combinations	7	0 (4)	0 (4)	1(3)	0 (3)	-	-	1	0
Sequences & Series	7	1(2)	0 (2)	1(1)	0 (1)	1(4)	0 (4)	3	0
TOTALS	**49**	**1**	**0**	**4**	**0**	**1**	**2**	**6**	**2**

Notes. The numbers in parentheses are the total numbers of items within the cell classified by thinking (process) skills and curricular content. - = no items were classified in the cell. n_i = number of items.

Table 10.

Biology DIF Items Organized by Unit Topic: Dichotomous Items

Unit Topic	Items	Male	Female
Nervous & Endocrine System	17	1	0
Reproductive Systems & Hormones	4	0	0
Differentiation & Development	4	0	1
Cell Division & Mendelian Genetics	18	1	1
Molecular Genetics	9	1	0
Population Genetics & Interaction	16	0	0
TOTALS	**56**	**3**	**2**

Figure 1.

Contingency Table for kth level of the Matching Variable

	Item Response Categories					
Group	y_1	y_2	y_3	...	y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	...	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	...	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	...	n_{+Tk}	n_{++k}